가속기 구조에서의 그래프 신경망 성능 특성 분석

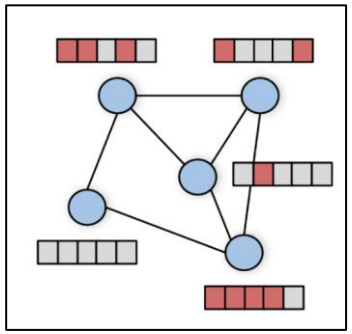
이훈종[1] 구건재[2]

고려대학교 컴퓨터학과

hunjong@korea.ac.kr[1] gunjaekoo@korea.ac.kr[2]

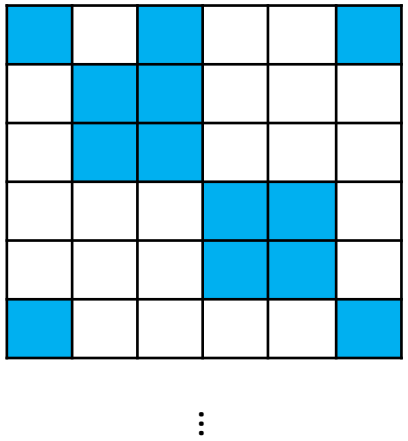
Sparsity in Graph Convolutional Network

Both Operands have high Sparsity !



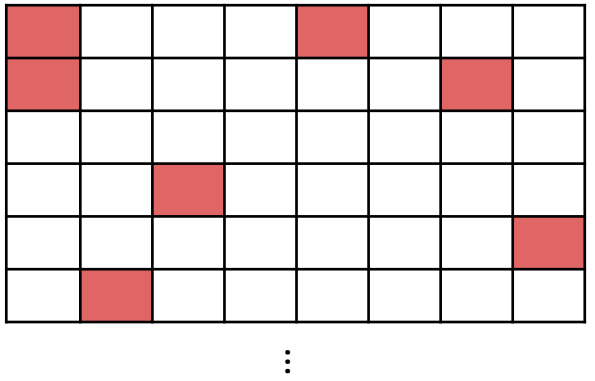
Aggregation[1]

=



Adjacency Matrix

*



Node Feature Matrix

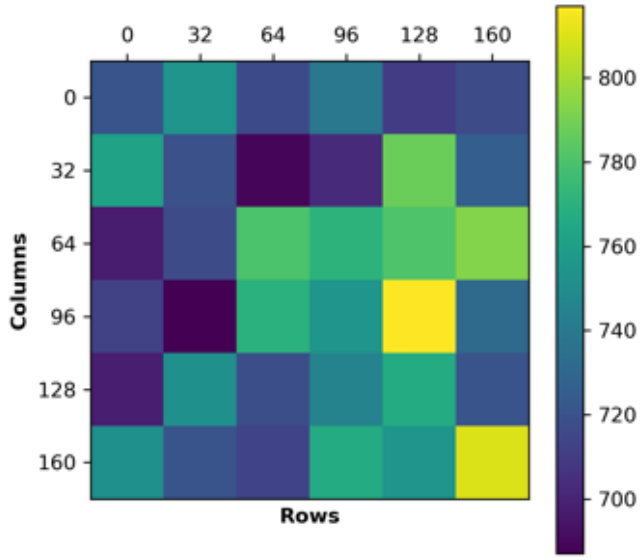
$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right)$$

[1] https://blog.twitter.com/engineering/en_us/topics/insights/2022/graph-machine-learning-with-missing-node-features

Convolutional Neural Network vs Graph Convolutional Network

GCN has higher sparsity than other DNN applications

CNN Sparsity

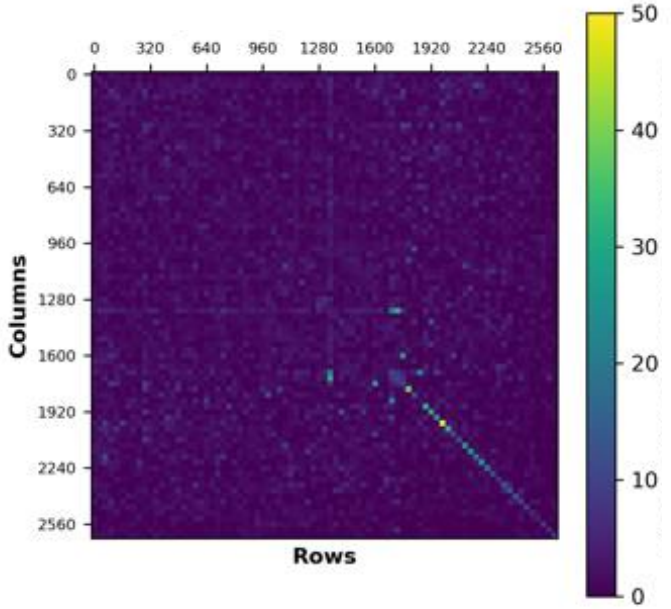


Input(Output) Activation of CNN

Average 50 ~ 60% Sparse

We can use SpGEMM Accelerator to overcome sparsity

GCN Sparsity



Adjacency Matrix in GCN

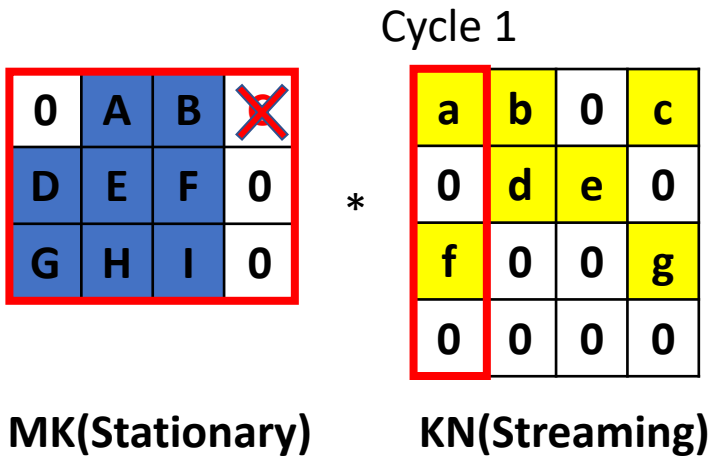
98% Sparse

Can we use SpGEMM Accelerator?

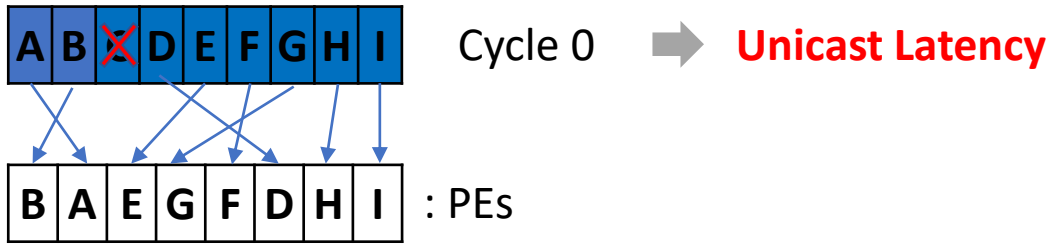
SIGMA

They use bitmap format to exploit structural sparsity and locality

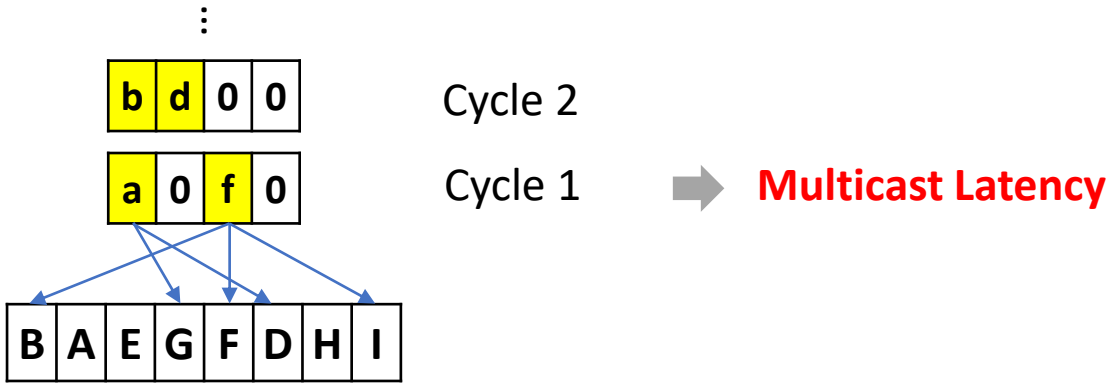
= But is it efficient in GCN?



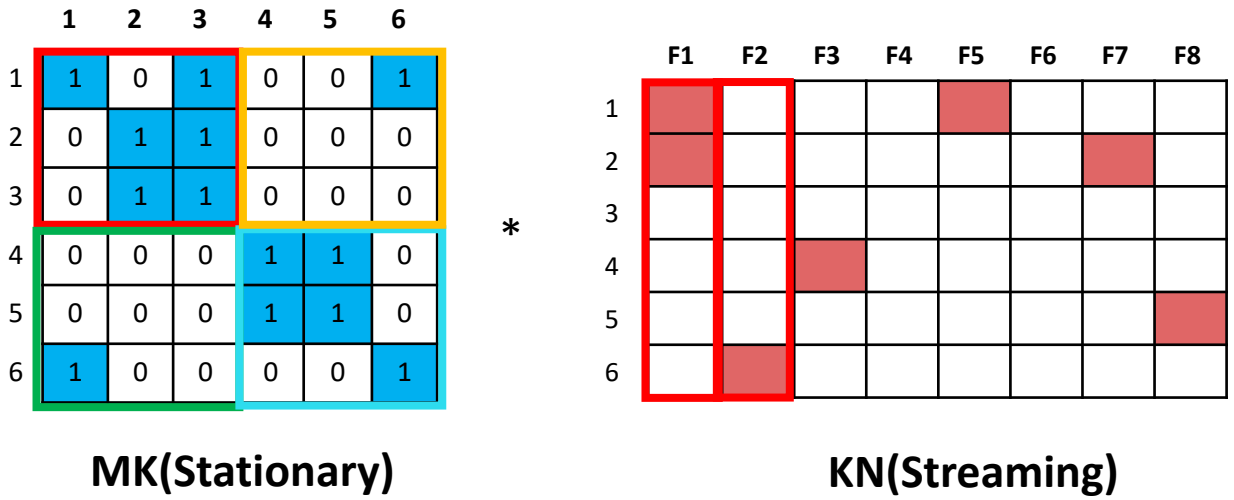
1. Unicast MK Matrix



2. Multicast KN Matrix

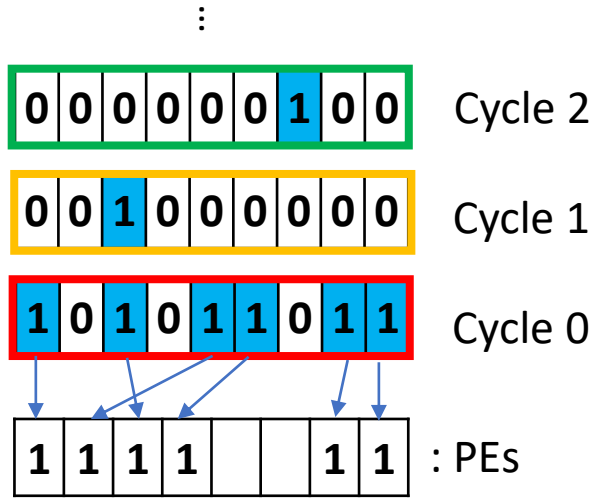


Bottleneck of SIGMA in GCN Aggregation Kernel



MK(Stationary) **KN(Streaming)**

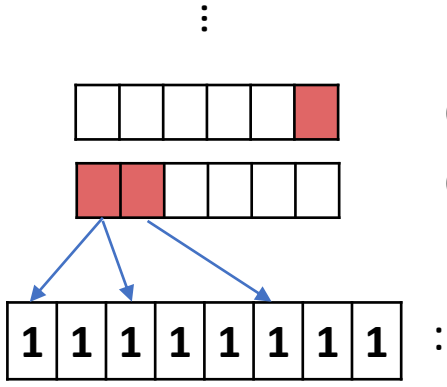
1. Unicast MK Matrix



Unicast Overheads

Increase Number of Unicast!

2. Multicast KN Matrix



Multicast Overheads

Bandwidth disturbed by zero-values

Evaluation Environment

We evaluate **1. Unicast Latency** **using highly sparse graph datasets**
2. Streaming Latency
3. Number of SRAM Read/Write

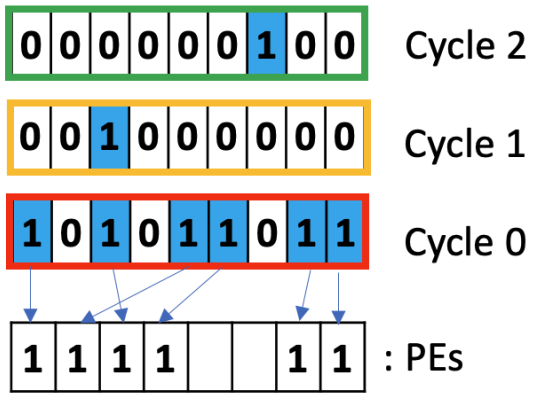
Architecture	Baseline	SIGMA
Dataflow	Output Stationary	MK_STA, KN_STR
Multiplier	256	256
Reduction Bandwidth	256	256
Distribution Bandwidth	32	256
Sparsity_Support	X	O

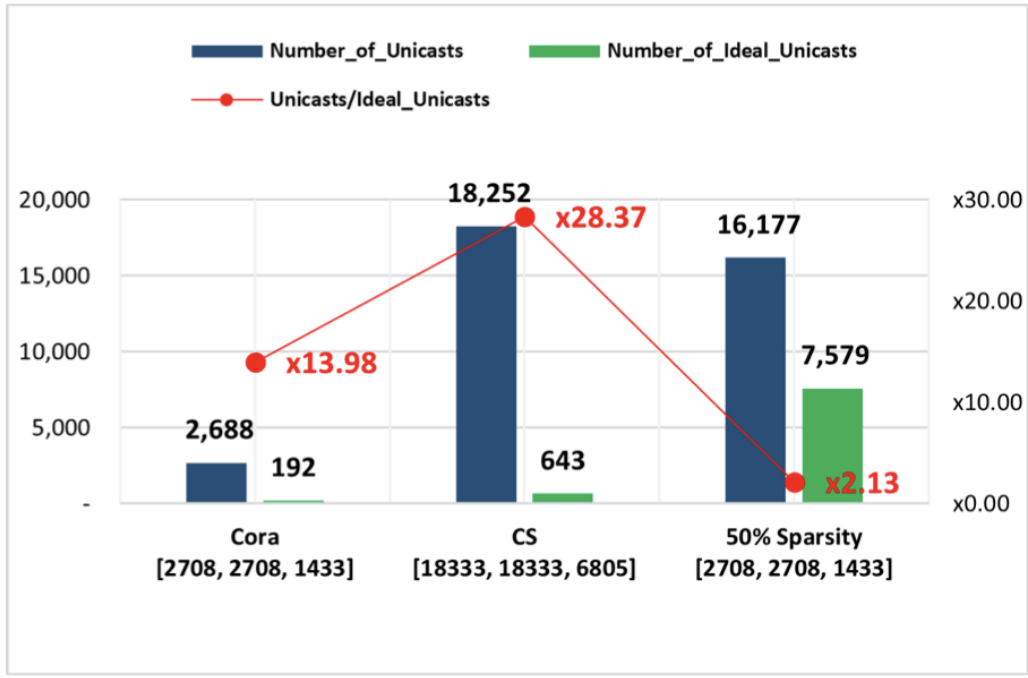
[STONNE Simulator Configuration]

Datasets	Cora	Couathor Physics	Coauthor CS	CitationFull DBLP	CitationFull PubMed
Adj_Density	0.1439%	0.0417%	0.0487%	0.0337%	0.0228%
Input_Density	1.268%	0.392%	0.875%	0.318%	10.022%

[Graph Dataset's Adjacency, Input Matrix Density]

Unicast Overheads in SIGMA

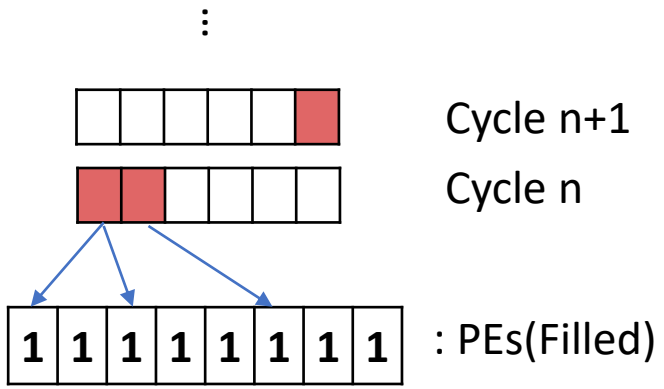


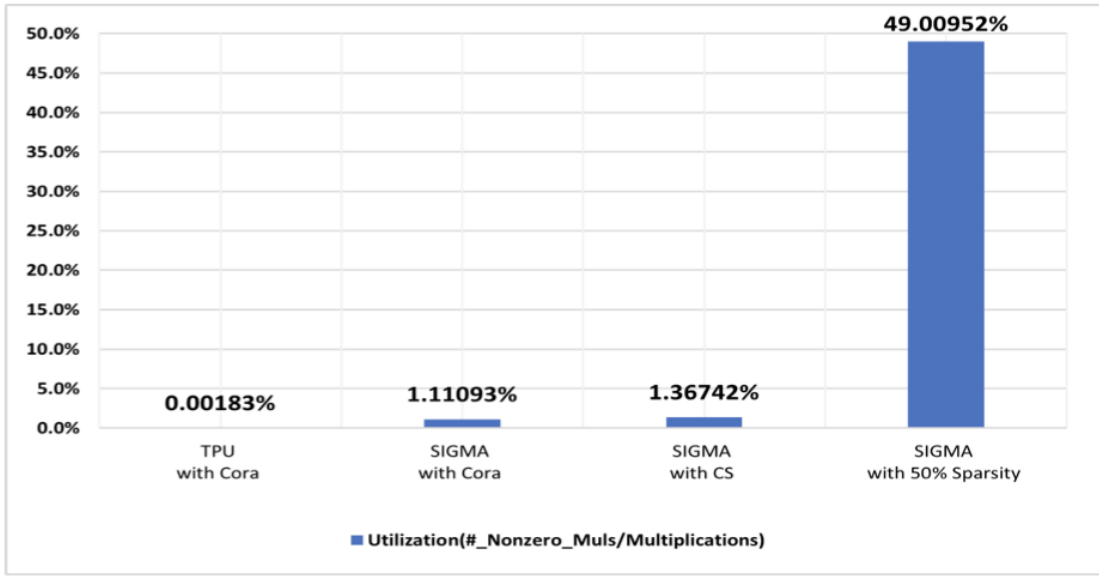


[Unicast Overhead in SIGMA(Loading Latency)]

Increased number of unicast due to SIGMA's structural gathering

PE(Processing Element) Underutilization

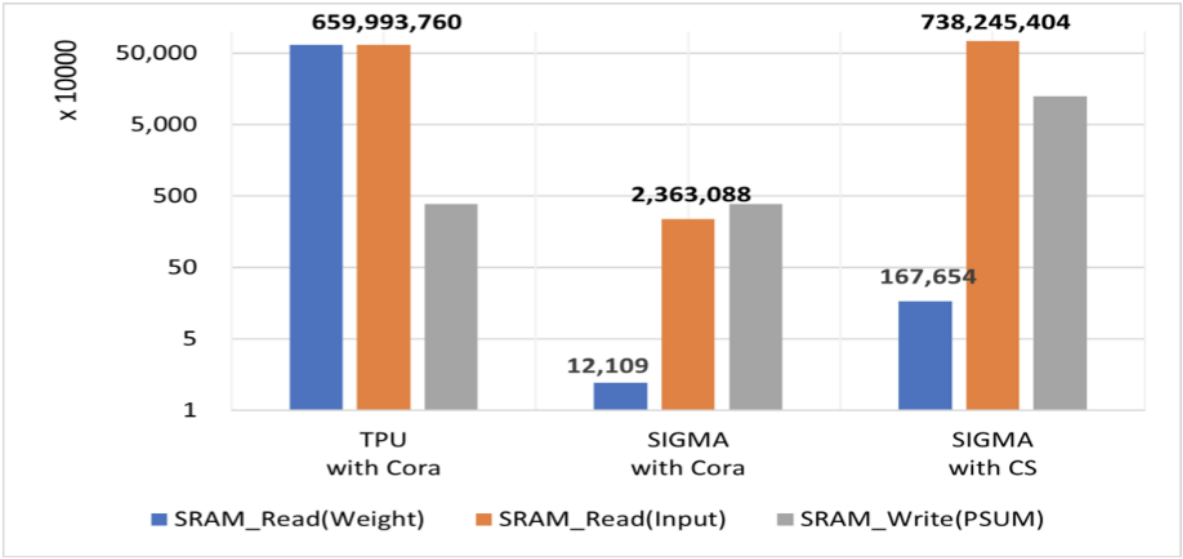




[PE Utilization]

Numerous zero-values disturb bandwidth and PE Utilization

SRAM Read/Write



[Number of SRAM Reads/Writes]

Needs more number of SRAM Reads due to transferring Numerous Zero-values